A Refined Mean Field Approximation

NICOLAS GAST, Inria, France
BENNY VAN HOUDT, University of Antwerp, Belgium

Mean field models are a popular means to approximate large and complex stochastic models that can be represented as N interacting objects. Recently it was shown that under very general conditions the steady-state expectation of any performance functional converges at rate $O(1/N)$ to its mean field approximation. In this paper we establish a result that expresses the constant associated with this $1/N$ term. This constant can be computed easily as it is expressed in terms of the Jacobian and Hessian of the drift in the fixed point and the solution of a single Lyapunov equation. This allows us to propose a refined mean field approximation. By considering a variety of applications, that include coupon collector, load balancing and bin packing problems, we illustrate that the proposed refined mean field approximation is significantly more accurate that the classic mean field approximation for small and moderate values of N: relative errors are often below 1% for systems with $N = 10$.

ACM Reference Format:
Nicolas Gast and Benny Van Houdt. 2017. A Refined Mean Field Approximation. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 2, Article 33 (December 2017), 28 pages. https://doi.org/10.1145/3152542

1 INTRODUCTION

Stochastic models have been used to assess the performance of computer (and many other) systems for many decades. As a direct analysis of large and complex stochastic models is often prohibitive, approximations methods to study their behavior have been devised. One very popular approximation method relies on mean field theory. Its widespread use can be explained by the relative ease involved to define and solve a mean field model in combination with its high accuracy for large systems. More specifically, many mean field models are specified by a simple set of ordinary differential or difference equations and in many cases convergence of the stochastic system towards the solution of the mean field model (over both finite and infinite time scales) has been established as the number of components N in the system tends to infinity, see [2, 3, 11, 15, 16, 18]. Moreover these mean field approximations may give rise to explicit expressions for the performance measures of interest (e.g., [21, 22, 27, 29]) or can be solved numerically without much effort (e.g., [10, 20, 28]).

Although mean field theory provides no guarantees with respect to the accuracy of a finite system of size N, there are many examples where the accuracy was demonstrated (using simulation) to be quite high even for systems of moderate size, e.g., for $N \approx 50$ or less. Nevertheless this accuracy very much depends on the exact parameter settings, for instance mean field models for load balancing systems are known to be quite inaccurate under high loads even for moderate sized systems.

The main contribution of this paper is to establish a result that applies to a broad class of mean field models, including density dependent population processes [15] and the class of discrete-time models of [2], and that provides a significantly more accurate approximation for finite N than the

Authors' addresses: Nicolas Gast, Inria, Univ. Grenoble Alpes, CNRS, LIG, Grenoble, F-38000, France, nicolas.gast@inria.fr; Benny Van Houdt, University of Antwerp, Middelheimlaan 1, Antwerp, B-2020, Belgium, benny.vanhoudt@uantwerpen.be.

2476-1249/2017/12-ART33 $15.00
https://doi.org/10.1145/3152542

	Coupon	Supermarket	Pull/push
Simulation ($N = 10$)	1.530	2.804	2.304
Refined mean field ($N = 10$)	1.517	2.751	2.295
Mean field ($N = \infty$)	1.250	2.353	1.636

Table 1. Illustration of the refined approximation accuracy. The columns correspond to the mean number of users with coupon i ($K = 5$, Table 2), the average queue length for the supermarket model ($\rho = 0.9$, Table 4) and for the pull/push model ($\rho = 0.9$, Table 7).

classical mean field approximation. When considering a mean field model $X^{(N)}$ described by the density-dependent population process of Kurtz, where $X_i^{(N)}(t)$ is the fraction of objects in state i at time t, the classical mean field approximation shows that if the corresponding ODE model has a unique attractor π, then

$$\lim_{N\to\infty} \mathbf{E}^{(N)}\left[\left\|X^{(N)} - \pi\right\|\right] = 0.$$

This shows in particular that for a continuous function h, we have:

$$\lim_{N\to\infty} \mathbf{E}^{(N)}\left[h(X^{(N)})\right] = h(\pi).$$

We establish conditions that show that for a function h, there exists a constant V_h such that

$$\mathbf{E}^{(N)}\left[h(X^{(N)})\right] = h(\pi) + \frac{V_h}{N} + o\left(\frac{1}{N}\right).$$

The constant V_h is expressed as a function of the Jacobian and the Hessian matrices of the drift (in π) and the solution of a single Lyapunov equation. As such it can be computed efficiently even when the number of possible states i of an object is large (less than one second for a 500-dimensional model). We also establish that under an additional smoothness condition, the $o(1/N)$ can be expressed as $V_h^{(2)}/N^2 + \cdots + V_h^{(k)}/N^k + o(1/N^k)$, but the computation of these constants is computationally intensive.

The refined mean field model that we propose consists in approximating $\mathbf{E}^{(N)}[h(X^{(N)})]$ by $h(\pi) + V_h/N$ and maintains many of the attractive features of the classic mean field approximation that resulted in its widespread use, but significantly improves upon its accuracy for finite N. We demonstrate that this refined approximation significantly improves the accuracy of the mean field approximation by using a variety of mean field models that include the coupon replication model [19, 20] and a number of load balancing schemes [21, 22, 29, 30]. In some special cases the constant V_h can be expressed in closed form, while for remaining cases the computation time for a numerical evaluation of V_h is negligible.

Our contributions are twofold:

- From a theoretical point of view, the main contribution is to prove the existence of the corrected term V_h and to provide an algorithmic way to compute it (Theorem 3.1). This proves that it is possible to define and compute the refined approximation $h(\pi) + V_h/N$.
- From a more practical perspective, we use a variety of examples that demonstrate that the refined approximation $h(\pi) + V_h/N$ is often highly accurate. To give a flavor of the accuracy of the refined approximation, we provide some examples in Table 1 picked from Sections 4, 5 and 6. This table illustrates that the refined approximation is very accurate, even for $N = 10$. Moreover, it is much more accurate than the classic mean field approximation.

Our goal when presenting many examples is to show that the refined approximation is accurate in many scenarios. Also, each of the examples developed in Sections 4–7 serves a particular purpose.

Section 4 shows that it is sometimes possible to obtain a closed form formula for the refined term V_h. Section 5 presents the supermarket model, which is one of the most widely studied mean field models. We also use this example to show that our refined approximation is accurate enough to quantify $1/N$-effects such as the impact of choosing with or without replacement among N servers. Section 6 shows that there are some cases for which the refined approximation is less than 1% from the actual value although the original mean field approximation underestimate the response time by a factor 2. Section 7 presents an example in which the quantity of interest is a blocking probability. All the previous examples are continuous-time examples that fall into the scope of density-dependent population process. Appendix A shows that it is possible to apply our results to discrete-time models as well.

Our results are obtained by using Stein's method (see [5] for a good introduction to Stein's method). The use of Stein's method to compute bounds for stationary distributions has been recently popularized in [4, 12]. This methodology has then been used in [31] and further extended in [9, 32] to establish the rate of convergence of stochastic processes to their mean field approximation, in light or heavy traffic. Our paper strongly relies on the methodology developed in those papers. In particular, it is shown in [9] that the mean field approximation is $1/N$-accurate. We improve this result as we are able, for a large class of models, to express the constant V_h by a simple formula that can be easily evaluated numerically. Note that the idea of using $h(\pi) + V_h/N$ as an improved approximation was already proposed in [9] for the two-choice model, where the constant V_h was estimated by simulation. The idea of improving classical approximations via Stein's method was also presented in [4] but, according to the authors, their computations seem hard to generalize to high dimensional settings. On the contrary, our method scales as the cube of the number of dimensions of the model which makes it applicable to models with hundreds of dimensions.

Roadmap. The rest of the paper is organized as follows. In Section 2 we introduce the model. We develop the theory in Section 3. We present several models in Sections 4 to 7 for which we compare the refined approximation with simulation. We conclude in Section 8.

2 MODEL AND NOTATIONS

Our result applies to many mean field models, including density dependent population processes [15] or the classical model of [2].

2.1 Density-Dependent Population Processes

In this paper, we consider a model that generalizes the notion of density-dependent population processes introduced by Kurtz in [15]. Consider a sequence of continuous-time Markov chains (CTMC) $(X^{(N)}(t))_N$. For each N, $X^{(N)}(t)$ evolves on a bounded subset $\mathcal{E}$ of $\mathbb{R}^d$. There exists a (finite or countable) set of vectors $\mathcal{L} \in \mathbb{R}^d$ and a set of functions $\beta_{\boldsymbol{\ell}}^{(N)} : \mathcal{E} \to \mathbb{R}^+$ such that $X^{(N)}(t)$ jumps from x to $x + \boldsymbol{\ell}/N$ at rate $N\beta_{\boldsymbol{\ell}}^{(N)}(x)$ for each $\boldsymbol{\ell} \in \mathcal{L}$.

We will refer to the chain $X^{(N)}(t)$ as the system of size N. For such a system, we define the drift $f^{(N)}$ as

$$f^{(N)}(x) = \sum_{\boldsymbol{\ell} \in \mathcal{L}} \boldsymbol{\ell} \beta_{\boldsymbol{\ell}}^{(N)}(x).$$

The drift corresponds to the expected variation of a chain $X^{(N)}(t)$:

$$f^{(N)}(x) = \lim_{dt \to 0} \frac{1}{dt} \mathbb{E}\left[X^{(N)}(t + dt) - X^{(N)}(t) | X^{(N)}(t) = x\right].$$

The classical notion of a density-dependent population process corresponds to the case where the transitions rates $\beta_{\boldsymbol{\ell}}^{(N)}(x)$ do not depend on N.

2.2 Scaling Assumptions

In the following, we will prove our result under the assumptions (A0)–(A4) below. Note that Assumption (A2) ensures that from any initial condition $x \in \mathcal{E}$, the ODE $\dot{x} = f(x)$ has a unique solution. We will denote by $\Phi_t x$ the value of this solution at time t:

$$\Phi_t x = x + \int_0^t f(\Phi_s x) ds.$$

The operator Φ is called the semi-group of the differential equation.

(A0) $\sup_{x\in\mathcal{E}, N\geq 1} \sum_{\boldsymbol{\ell}\in\mathcal{L}} \beta_{\boldsymbol{\ell}}^{(N)}(x)\|\boldsymbol{\ell}\|^2 < \infty$.

(A1) There exist two functions $f : \mathcal{E} \to \mathcal{E}$ and $\tilde{f} : \mathcal{E} \to \mathcal{E}$ such that, uniformly in x:

$$f^{(N)}(x) = f(x) + \frac{1}{N}\tilde{f}(x) + O(\frac{1}{N^2}).$$

(A2) f is twice-differentiable and its second derivative is uniformly continuous and bounded.

(A3) The ODE $\dot{x} = f(x)$ has a unique attractor π and this attractor is exponentially stable, *i.e.* there exists $a, b > 0$ such that for all $x \in \mathcal{E}$:

$$\|\Phi_t x - \pi\| \leq a e^{-bt}.$$

(A4) For each N, the stochastic process has a unique stationary distribution. Note that for a function $h : \mathcal{E} \to \mathbb{R}$, we denote by $\mathbf{E}^{(N)}[h(X^{(N)})]$ the expected value of $h(X^{(N)})$ in steady-state.

The first assumption (A0) is mostly technical and is needed so that the variance of the stochastic process does not explode in finite time. The second assumption is always true if the rate functions $\beta_{\boldsymbol{\ell}}$ do not depend on N (which is the case for the classical model of Kurtz [15]). It is needed to have a $O(1/N)$ convergence term and will be commented later.

2.3 Discrete-Time Model of [2]

In this paper, we mainly focus on mean field interacting models that can be modeled as density-dependent population processes. Another type of model that is popular in the performance evaluation community is the mean field interaction model of [2]. In this section, we briefly sketch how to adapt our result in this case.

The authors of [2] consider a discrete-time model with N objects, where $S_n^N(t) \in \{1 \dots d\}$ denotes the state of the object n at time slot t. The objects are assumed to be interchangeable, meaning that the empirical measure $X^{(N)}$ is assumed to be a discrete-time Markov chain, where $X_i^{(N)}(t)$ is the proportion of objects in state i at time t, defined as

$$X_i^{(N)}(t) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_{\{S_n^N(t)=i\}}.$$

The main assumption of [2] is that there exists a function $\varepsilon(N)$, called the intensity of the stochastic process, that vanishes as N goes to infinity (*i.e.*, $\lim_{N\to\infty} \varepsilon(N) = 0$) and such that

$$\lim_{N\to\infty} \varepsilon(N)\mathbb{E}\left[X^{(N)}(t+1) - X^{(N)}(t) \mid X^{(N)}(t) = x\right] = f(x),$$

and such that the number of objects that change state during one time slot is bounded by $\varepsilon(N)$. It is then shown in [2, Theorem 1] that, the stochastic process $X^{(N)}(t/\varepsilon(N))$ converges in probability to a deterministic quantity $\Phi_t x$ at rate $\sqrt{\varepsilon(N)}$, where $t \mapsto \Phi_t x$ is the solution of the differential equation starting in x at time 0.

These discrete-time Markov chains can be transformed into continuous time Markov chains by assuming that each time-slot lasts for an exponential amount of time with mean $1/\varepsilon(N)$. This new chain has the same stationary distribution as the original discrete-time chain. Hence, as noted in [3, Section 6] it is possible to transform most of the discrete-time models of this type into density-dependent population processes for which we can apply Theorem 3.1 to show that

$$\mathbf{E}^{(N)}\left[h(X^{(N)})\right] = h(\pi) + \varepsilon(N)V_h + o(\varepsilon(N)).$$

We present an example in Appendix A.

3 MAIN RESULTS

3.1 Computation of the Constant

We consider a mean field model that has a unique fixed point π that is a global attractor. Let Q be the positive definite symmetric matrix whose (n,m)-th entry is equal to:

$$Q_{n,m} = \sum_{\boldsymbol{\ell}\in\mathcal{L}} \ell_n \ell_m \beta_{\boldsymbol{\ell}}(\pi).$$

We denote by A the first derivative of f in π (*i.e.* the Jacobian in π) and by B the second derivative of f, that is, for $i,j,k_1,k_2 \in \{1,\dots,d\}$:

$$A_{i,j} = \frac{\partial f_i}{\partial x_j}(\pi), \qquad (B_j)_{k_1,k_2} = \frac{\partial^2 f_j}{\partial x_{k_1}\partial x_{k_2}}(\pi).$$

As indicated by Assumption (A3), we assume that the ODE system $\dot{x} = f(x)$ is exponentially stable, which implies that the matrix A is a Hurwitz matrix (meaning the real part of any eigenvalue of A is negative) [14]. Therefore, by [26], the Lyapunov equation $AX + XA^T + Q = 0$ has a unique solution. We denote by W the unique symmetric matrix (as Q is symmetric) that satisfies

$$AW + WA^T + Q = 0. \tag{1}$$

Last, we define by C the vector :

$$C = \tilde{f}(\pi) = \lim_{N\to\infty} N(f^N(\pi) - f(\pi)).$$

The vector C is equal to zero if the transition rates do not depend on N (this is the case in all of our examples except for the coupon collector application in Section 4.1 and the supermarket without replacement of Section 5.3).

We are now ready to state the main theorem of the paper, whose proof is postponed to Section 3.4.

Theorem 3.1. *Assume that the model satisfies (A0–A4). Let $h : \mathcal{E} \to \mathbb{R}$ be a twice-differentiable function that has a uniformly continuous second derivative. Then,*

$$\lim_{N\to\infty} N\left(\mathbf{E}^{(N)}\left[h(X^{(N)})\right] - h(\pi)\right) = \sum_i \frac{\partial h}{\partial x_i}(\pi)V_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2 h}{\partial x_i \partial x_j}(\pi)W_{ij}, \tag{2}$$

where the matrices A, C and W are defined above and V_i is equal to:

$$V_i = -\sum_j (A^{-1})_{i,j}\left[C_j + \frac{1}{2}\sum_{k_1,k_2}(B_j)_{k_1,k_2}W_{k_1,k_2}\right]. \tag{3}$$

In the following, we will mainly use this result to develop a refined approximation of the expectation of $X_i^{(N)}$, which is asymptotically equal to $\pi_i + V_i/N + o(1/N)$. We also note that this result can also be used to compute the co-variance matrix of the vector $\sqrt{N}(X^{(N)} - \pi)$. Both of these results are a direct consequence of Theorem 3.1 and are summarized in the following corollary.

Corollary 1. *Assume that the model satisfies (A0–A4). Then, using the same notations as in Theorem 3.1, we have:*

(i) *For any coordinate i,*

$$\lim_{N\to\infty} N\left(\mathbf{E}^{(N)}\left[X_i^{(N)}\right] - \pi_i\right) = V_i;$$

(ii) *For any coordinates i, j, the covariance satisfies:*

$$\lim_{N\to\infty} N\mathrm{cov}\left(X_i^{(N)}, X_j^{(N)}\right) = W_{ij}.$$

Proof. The proof of *(i)* is a direct application of Theorem 3.1 with the function $h(x) = x_i$. For *(ii)*, as π_i is deterministic, we have

$$\begin{aligned} N\mathrm{cov}(X_i, X_j) &= N\mathrm{cov}(X_i - \pi_i, X_j - \pi_j) \\ &= N\mathbf{E}^{(N)}\left[(X_i - \pi_i)(X_j - \pi_j)\right] - N(\mathbf{E}^{(N)}[X_i] - \pi_i)(\mathbf{E}^{(N)}[X_j] - \pi_j) \end{aligned}$$

applying Theorem 3.1 with $h(x) = (x_i - \pi_i)(x_j - \pi_j)$ shows that the first term converges to W_{ij} and the second is of order $O(1/N)$. □

The above result shows that $\mathbb{E}[X_i^{(N)}]$ is of order $\pi_i + V_i/N$ plus a $o(1/N)$-term. In fact, a natural extension of Theorem 3.1 shows that, if the drift f is differentiable 3 times, then this higher-order term is a $O(1/N^2)$-term, and in that case we have :

$$\mathbf{E}^{(N)}\left[X_i^{(N)}\right] = \pi_i + \frac{1}{N}V_i + O\left(\frac{1}{N^2}\right).$$

This result can be readily extended to higher-order differentiable drifts.

Theorem 3.2. *Assume that the model satisfies Assumptions (A0–A4). If in addition the drift is continuously differentiable k times, then for each coordinate i, there exists a series of constants $V_i^{(2)} \dots V_i^{(k-1)}$ such that*

$$\mathbf{E}^{(N)}\left[X_i^{(N)}\right] = \pi_i + \frac{1}{N}V_i + \frac{1}{N^2}V_i^{(2)} \cdots + \frac{1}{N^{k-1}}V_i^{(k-1)} + o\left(\frac{1}{N^{k-1}}\right),$$

where V_i is defined as in Theorem 3.1.

The result is an (almost direct) consequence of Lemma 3.5 that can be used to show that similarly to V_i/N, the constant $V_i^{(j)}$ can be expressed as a function of some integral of the flow Φ_t and its derivative in π up to the jth one. However, apart from particular cases, the computational cost of a numerical evaluation of these constants grows quickly with k. Hence, we believe that from a practical point of view, studying beyond $k = 2$ is not feasible. We provide some details about it in Appendix B.

3.2 Algorithmic Considerations

In some cases an explicit expression for the matrix W and the vector V in Theorem 3.1 can be obtained (*e.g.*, see Section 4). However, in general the constant V_i must be computed by a numerical procedure that consists of the following steps:

(1) Computation of the fixed point π and definition of Q.
(2) Computation of the first and second derivative A and B.
(3) Computation of W by solving the Lyapunov Equation (1).
(4) Application of (3) to compute the vector V.

This procedure is a direct application of Theorem 3.1 and can be easily automated[1]. We further note that most of the conditions needed to apply the theorem can be checked automatically, at the exception of Assumption (A3): proving that a dynamical system has a unique attractor is undecidable [24].

The time-complexity of the numerical procedure is essentially $O(d^3)$, where d is the dimension of the mean field model. The main step in the computation is the determination of the solution W to the Lyapunov equation $AX + XA^T + Q = 0$. This solution can be computed in $O(d^3)$ time using the Bartels-Stewart algorithm [1] (which is implemented by the MATLAB function `lyap` and by the function `scipy.linalg.solve_lyapunov` of the `scipy` package of `python` [13]). The other steps of the algorithm are less costly. Because we assume that there exists a unique fixed point that is exponentially stable, the fixed point can be easily computed by numerically integrating the ODE and/or using a Newton-Raphson iteration close to π. Once the fixed point π has been computed, the matrices A and B can be computed by using a symbolic toolbox. The computation of V_i for all i can also be done in $O(d^3)$.

In our numerical examples, we tested up to $d = 500$ for which the computation time was less than one second. We further note that in many cases the fixed point π can be obtained in closed form. Also, in some particular cases the entries of $(-A)^{-1}$ can also be expressed in explicit form (*e.g.*, see Section 5). Even in the cases when W and A^{-1} have to be computed numerically, the numerical computation of these constants is several orders of magnitude faster than simulation.

3.3 Extensions to Infinite-Dimensional Models and Unbounded State-Spaces

To simplify the exposition of the theory, we mostly restrict our attention to population processes that evolve in a subset of $\mathbb{R}^d$. This excludes infinite-dimensional population models that arise naturally in queuing networks when the queues are unbounded like in the examples of Section 5 and Section 6 for which the state-space is a subset of $[0,1]^{\mathbb{Z}^+}$. The main reason to work with finite-dimensional models is that in many cases, we evaluate numerically the constants V and W by using numerical methods to invert A and to solve the Lyapunov equation (1). This requires the use of finite matrices. In our numerical examples, the only infinite-dimensional models are the supermarket model of Section 5 and the push-pull model of Section 6. To obtain our numerical results, we truncated the models to models of dimension $d = 40$ for the supermarket model and $d = 500$ for the push-pull model. Higher values of d lead to the same refined constants V_i up to the floating point errors.

From a theoretical point of view, Theorem 3.1 can be easily generalized to a Banach state space. One key issue, however, is that Assumption (A3) will in general not hold if the state-space $\mathcal{E}$ is not bounded, which is in general the case when $\mathcal{E}$ is infinite-dimensional (in an unbounded space with bounded drift, the attractor cannot attract the trajectories in finite time). A possible way to overcome this difficulty is to replace Assumption (A3) by the weaker Assumption (A5): which is to assume that (A3) holds on a bounded subset $\mathcal{B} \subset \mathcal{E}$ and that the sequence of stationary measures of the stochastic processes concentrates on $\mathcal{B}$ at rate $1/N^3$:

(A5) There exists a bounded set $\mathcal{B} \subset \mathcal{E}$ such that:

(i) The ODE $\dot{x} = f(x)$ has a unique attractor π and there exists $a, b > 0$ such that *for all* $x \in \mathcal{B}$:

$$\|\Phi_t x - \pi\| \leq a e^{-bt}.$$

(ii) There exists a constant $B > 0$ such that in steady-state: $\mathbf{P}[X^{(N)} \notin \mathcal{B}] \leq B/N^3$.

Note that the only difference between Assumption (A3) and Assumption (A5)–(i) is that the latter holds for $x \in \mathcal{B}$ (and not for all $x \in \mathcal{E}$ as in Assumption (A3)).

[1] A prototype of a tool that implements this algorithm is available at https://github.com/ngast/rmf_tool/.

As for (A3), establishing (A5) is model-specific. For the supermarket and the pull-push model, Assumption (A5)–(i) was established in [22] and [21]. Assumption (A5)–(ii) can be proved by using a coupling argument that shows that the processes is bounded by a system of N independent $M/M/1$ queues, see [9, Section 6.4] for more details. This coupling argument seems to be easily generalizable to many models of resource allocation strategies.

We then obtain the following theorem:

THEOREM 3.3. *Assume that the state-space $\mathcal{E}$ is a Banach space and that assumptions (A0,A1,A2,A4,A5) hold. Then, for any twice differentiable bounded function h, Theorem 3.1 continues to hold and the constants V and W satisfy the same equations.*

SKETCH OF PROOF. Let $\mathcal{B}$ be the bounded set of Assumption (A5). We proceed as for the proof of [9, Theorem 3.2]. First, it should be noted that one may assume without loss of generality that h equals $h(\pi)$ outside a bounded set. Then, as shown in [9, Equation (10)], we have:

$$\begin{aligned} N(\mathbb{E}\left[h(X^{(N)})\right] - h(\pi)) =& N\mathbf{E}^{(N)}\left[(\Lambda - L^{(N)})Gh(X^{(N)})\mathbf{1}_{\{X^{(N)}\notin\mathcal{B}\}}\right] \\ &+ N\mathbf{E}^{(N)}\left[(\Lambda - L^{(N)})Gh(X^{(N)})\mathbf{1}_{\{X^{(N)}\in\mathcal{B}\}}\right]. \end{aligned}$$

If h is bounded, then by assumption (A5)–(ii), the right-hand side of the first line is of order $O(1/N^2)$ (this can be shown by adapting the proof of [9, Lemma 6.4]). Hence, the result follows if one can show that the second term converges to Equation (2).

To show that, one needs to generalize Lemma 3.4, 3.5 and 3.6 to a Banach space. The essential difference lies in the fact that the h-th derivative of function $x \mapsto \int_0^\infty \Phi_t x dt$ is bounded on any bounded set, but not necessarily on the whole state-space $\mathcal{E}$ (see [9, Lemma 6.3]). This is why we need (A5). Apart from that, all the other arguments continue to hold – for instance, the uniqueness of the solution of the Lyapunov equation comes from the link between the exponential stability of the ODE and the Lyapunov equation in a Banach space that is established in [7, 25]. □

3.4 Proof of Theorem 3.1

In what follows, for a function $h : \mathcal{E} \to \mathbb{R}$, we will denote by Dh the derivative of h (if it exists) and by D^2h its second derivative. The notations $Dh(x)\cdot\boldsymbol{\ell}$ and $D^2h(x)\cdot(\boldsymbol{\ell},\boldsymbol{\ell})$ stand respectively for $\sum_i(\partial h(x)/\partial x_i)\boldsymbol{\ell}_i$ and $\sum_{i,j}(\partial^2 h(x)/(\partial x_i\partial x_j))\boldsymbol{\ell}_i\boldsymbol{\ell}_j$. For a multivariate function $g : \mathcal{E} \to \mathcal{E}$, we also use the notations Dg and D^2g for the first and second derivative. $(Dg)_{ij}$ denotes $\partial g_i/\partial x_j$ and $(D^2g)_{i,k_1,k_2}$ denotes $\partial^2 g_i/(\partial x_{k_1}\partial x_{k_2})$.

PROOF OF THEOREM 3.1. We denote by $L^{(N)}$ the generator of the chain $X^{(N)}(t)$ and by Λ the generator of the ODE. They associate to a differentiable function $h : \mathcal{E} \to \mathbb{R}$ the functions $L^{(N)}h : \mathcal{E} \to \mathbb{R}$ and $\Lambda h : \mathcal{E} \to \mathbb{R}$ defined by

$$\begin{aligned} (L^{(N)}h)(x) &= \sum_{\boldsymbol{\ell}\in\mathcal{L}} N\beta_{\boldsymbol{\ell}}^{(N)}(x)(h(x+\frac{\boldsymbol{\ell}}{N}) - h(x)), \\ (\Lambda h)(x) &= Dh(x)\cdot f(x) = \sum_{\boldsymbol{\ell}\in\mathcal{L}} \beta_{\boldsymbol{\ell}}(x) Dh(x)\cdot\boldsymbol{\ell}. \end{aligned}$$

The proof can be divided in three steps. The first step is to use Stein's method and the ideas developed in [9, 31] to show that, in steady-state, we have

$$\mathbf{E}^{(N)}\left[h(X^{(N)})\right] - h(\pi) = \mathbf{E}^{(N)}\left[(L^{(N)} - \Lambda)\int_0^\infty (h(\Phi_s X^{(N)}) - h(\pi))ds\right]. \tag{4}$$

This is a direct consequence of Equation (10) of [9].

The second step is to compare the generators $L^{(N)}$ and Λ to relate the above equation and the differential of the function $G : x \mapsto \int_0^\infty (\Phi_s x - \pi)ds$, which is well-defined by (A3). This is done in Lemma 3.4.

Using perturbation theory, we then show in Lemma 3.5 that the function $G_h : x \mapsto \int_0^\infty (h(\Phi_s x) - h(\pi))ds$, which is the solution of the Poisson equation $(\Lambda G_h)(x) = h(\pi) - h(x)$ [9], satisfies the assumptions of Lemma 3.4. This shows that:

$$\lim_{N\to\infty} N\mathbf{E}^{(N)}\left[(L^{(N)} - \Lambda)\int_0^\infty (h(\Phi_s X^{(N)}) - h(\pi))ds\right] = DG_h(\pi)\cdot C + \frac{1}{2}\sum_{\boldsymbol{\ell}\in\mathcal{L}} \beta_{\boldsymbol{\ell}}(\pi) D^2 G_h(\pi)\cdot(\boldsymbol{\ell},\boldsymbol{\ell}).$$

The proof is completed by Lemma 3.6 in which we derive the general formula presented in Equation (3). □

Lemma 3.4. *Let g be a twice differentiable function that has a uniformly continuous second derivative that is bounded. Then, for $x \in \mathcal{E}$:*

$$\lim_{N\to\infty} N(L^{(N)} - \Lambda)gx = Dg(x)\cdot\tilde{f}(x) + \sum_{\boldsymbol{\ell}\in\mathcal{L}} \beta_{\boldsymbol{\ell}}(x) D^2 g(x)\cdot(\boldsymbol{\ell},\boldsymbol{\ell}).$$

Moreover the convergence is uniform in x.

Proof. By definitions of the generators, we have

$$\begin{aligned}(L^{(N)} - \Lambda)gx &= \sum_{\boldsymbol{\ell}\in\mathcal{L}} N\beta_{\boldsymbol{\ell}}^{(N)}(x)\left(g(x+\frac{\boldsymbol{\ell}}{N}) - g(x)\right) - Dg(x)\cdot f(x)\\ &= \sum_{\boldsymbol{\ell}\in\mathcal{L}} N\left(\beta_{\boldsymbol{\ell}}^{(N)}(x) - \beta_{\boldsymbol{\ell}}(x)\right)\left(g(x+\frac{\boldsymbol{\ell}}{N}) - g(x)\right)\\ &\quad + \sum_{\boldsymbol{\ell}\in\mathcal{L}} \beta_{\boldsymbol{\ell}}(x)\left(Ng(x+\frac{\boldsymbol{\ell}}{N}) - Ng(x) - Dg(x)\cdot\boldsymbol{\ell}\right). \qquad (5)\end{aligned}$$

We first bound the first summation in Equation (5). As g has a bounded second derivative, there exists a constant $c > 0$ such that $\|g(x+\frac{\boldsymbol{\ell}}{N}) - g(x) - Dg(x)\cdot\boldsymbol{\ell}/N\| \le c\|\boldsymbol{\ell}\|^2/N^2$. This shows that (uniformly in x), the first summation in Equation (5) can be simplified to:

$$\begin{aligned}\sum_{\boldsymbol{\ell}\in\mathcal{L}} N\left(\beta_{\boldsymbol{\ell}}^{(N)}(x) - \beta_{\boldsymbol{\ell}}(x)\right)\left(g(x+\frac{\boldsymbol{\ell}}{N}) - g(x)\right) &= \sum_{\boldsymbol{\ell}\in\mathcal{L}} N\left(\beta_{\boldsymbol{\ell}}^{(N)}(x) - \beta_{\boldsymbol{\ell}}(x)\right)\left(Dg(x)\cdot\frac{\boldsymbol{\ell}}{N} + O\left(\frac{\|\boldsymbol{\ell}\|^2}{N^2}\right)\right)\\ &= (Dg(x) + O\left(\frac{1}{N}\right))\cdot(f^N(x) - f(x)).\end{aligned}$$

which is equal to $Dg(x)\cdot\tilde{f}(x)/N + O(1/N^2)$ by Assumption (A1).

For the second summation in Equation (5), as $D^2 g$ is uniformly continuous in x, the quantity $N^2 g(x+\frac{\boldsymbol{\ell}}{N}) - N^2 g(x) - NDg(x)\cdot\boldsymbol{\ell}$ converges (uniformly in x) to $\frac{1}{2}D^2 g(x)\cdot(\boldsymbol{\ell},\boldsymbol{\ell})$. As a consequence, the second summation in Equation (5) is equal to $\frac{1}{2}\sum_{\boldsymbol{\ell}\in\mathcal{L}} \beta_{\boldsymbol{\ell}}(x) D^2 g(x)\cdot(\boldsymbol{\ell},\boldsymbol{\ell})/N + o(1/N)$. □

Lemma 3.5. *Let $(\Phi_t) : x \mapsto \Phi_t x$ be the flow associated to the drift $f : \mathcal{E} \to \mathbb{R}$ and $k \ge 0$. Assume that f and Φ are k-times differentiable with uniformly continuous derivatives. Assume that the ODE $\dot{x} = f(x)$ has a unique exponentially stable attractor π (Assumption (A3)). Let $h : \mathcal{E} \to \mathbb{R}$ be a k-times differentiable function that has a uniformly continuous kth derivative. Then, the function $G_h : x \mapsto \int_0^\infty (h(\Phi_s x) - h(\pi))ds$ is k-times continuously differentiable. Its kth derivative is bounded and equal to $\int_0^\infty D^k(h\circ\Phi_s)(x)ds$.*

Proof. According to [9, Lemma 6.3], there exists two constants $b > 0$ and $\delta > 0$ such that for all t: $\|D\Phi_t\| \leq be^{-\delta t}$. According to [8, Lemma C.1], this implies that for any $i \leq k+1$, the ith differentiable of Φ_t exists and there exists a constant M_i such that for all t: $\|D^i\Phi_t\| \leq M_i e^{-\delta t}$.

Let $G : x \mapsto \int_0^\infty (\Phi_s x - \pi)ds$. As $D^i\Phi_t$ converges exponentially fast to 0, the function G is i-times differentiable and its ith differentiable D^iG is equal to

$$D^iG(x) = \int_0^\infty D^i\Phi_t(x)ds.$$

By using the chain rule, the kth derivative of G_h can be expressed as a function of the derivatives of order 1 to k of G and h (this is a multivariate generalization of the *Faà di Bruno* formula). □

Lemma 3.6. *The following holds:*

(i) *The matrix equation $AX + XA^T + Q = 0$ has a unique solution, that we denote by W. Moreover, for k_1, k_2, we have :*

$$\sum_{\boldsymbol{\ell}\in\mathcal{L}} \beta_{\boldsymbol{\ell}}(\pi) \sum_{n,m} \ell_n \ell_m \int_{s=0}^\infty (e^{As})_{k_1,n}(e^{As})_{k_2,m}ds = W_{k_1,k_2}. \tag{6}$$

(ii) *The derivative of G satisfies:*

$$(DG(\pi))_{i,j} = (-A)^{-1}_{i,j}, \tag{7}$$

(iii) *The second derivative of G_h at π satisfies:*

$$\sum_{\boldsymbol{\ell}\in\mathcal{L}} \beta_{\boldsymbol{\ell}}(\pi) D^2G_h(\pi)\cdot(\boldsymbol{\ell},\boldsymbol{\ell}) = \sum_{i,j} \frac{\partial^2 h}{\partial x_i \partial x_j}(\pi) W_{i,j} + \sum_i \frac{\partial h}{\partial x_i}(\pi) \sum_j (-A)^{-1}_{i,j} \sum_{k_1,k_2} (B_j)_{k_1,k_2} W_{k_1,k_2}. \tag{8}$$

Proof. *(i)* – By definition of $Q_{n,m}$, Q is a positive definite matrix. As A is Hurwitz, there exists a unique positive definite matrix W that satisfies $AW + WA^T + Q = 0$ [26]. Moreover, this matrix W satisfies:

$$\sum_{n,m} \int_{s=0}^\infty (e^{As})_{k_1,n} Q_{n,m} (e^{A^T s})_{m,k_2} ds = W_{k_1,k_2}. \tag{9}$$

By definition, $Q_{n,m} = \sum_{\boldsymbol{\ell}\in\mathcal{L}} \beta_{\boldsymbol{\ell}}(\pi)\ell_n\ell_m$. This implies that Equation (6) is equivalent to (9).

(ii) – By the chain rule the derivative with respect to t of the Jacobian $D(\Phi_t)$ of the map $\Phi_t : x \to \Phi_t x$ is given by

$$\frac{d}{dt}(D(\Phi_t)(x))_{i,j} = \frac{\partial}{\partial x_j}(f_i \circ \Phi_t)(x) = \sum_u (Df(\Phi_t x))_{i,u}(D(\Phi_t)(x))_{u,j}.$$

As $\Phi_t\pi = \pi$, we have $Df(\Phi_t\pi)_{i,u} = A_{i,u}$ and therefore $\frac{d}{dt}(D(\Phi_t)(\pi)) = AD(\Phi_t)(\pi)$. This implies that

$$(D(\Phi_t)(\pi))_{i,j} = (e^{At})_{i,j}, \tag{10}$$

which yields (7) as $\int_{t=0}^\infty (e^{At})_{i,j}dt = (-A)^{-1}_{i,j}$ due to assumption (A3).

(iii) – By the chain rule, the derivative of the Hessian $D^2(\Phi_t)$ of the map Φ_t is given by

$$\begin{aligned}\frac{d}{dt}(D^2(\Phi_t)(x))_{j,nm} = &\sum_u (Df(\Phi_t x))_{j,u}(D^2(\Phi_t)(x))_{u,nm} \\ &+ \sum_{k_1,k_2} (D^2 f(\Phi_t x))_{j,k_1k_2}(D(\Phi_t)(x))_{k_1,n}(D(\Phi_t)(x))_{k_2,m}.\end{aligned}$$

Applying the above equality at π implies that

$$\frac{d}{dt}(D^2(\Phi_t)(\pi))_{j,nm} = \sum_u A_{j,u}(D^2(\Phi_t)(\pi))_{u,nm} + \sum_{k_1,k_2} B_{j,k_1k_2}(e^{At})_{k_1,n}(e^{At})_{k_2,m}.$$

The variation of constants method shows that

$$(D^2(\Phi_t)(\pi))_{i,nm} = \sum_j \int_{s=0}^{t} (e^{A(t-s)})_{i,j} \sum_{k_1,k_2} B_{j,k_1k_2}(e^{At})_{k_1,n}(e^{At})_{k_2,m} ds.$$

By switching the order of integration this implies that

$$\int_0^\infty (D^2\Phi_t(\pi))_{i,nm} = \sum_j (-A^{-1})_{i,j} \sum_{k_1,k_2} B_{j,k_1k_2} \int_{s=0}^{\infty} (e^{As})_{k_1,n}(e^{As})_{k_2,m} ds. \tag{11}$$

By definition of G_h, we have:

$$\begin{aligned}(D^2(G_h)(\pi))_{nm} = &\sum_i \frac{\partial h}{\partial x_i}(\pi) \int_0^\infty (D^2(\Phi_t)(\pi))_{i,nm} dt \\ &+ \sum_{i,j} \frac{\partial^2 h}{\partial x_i \partial x_j}(\pi) \int_0^\infty (D(\Phi_t)(\pi))_{i,n}(D(\Phi_t)(\pi))_{j,m} dt.\end{aligned}$$

The equality in (8) now follows by combining Equation (6), (10) and (11) with the above equation. □

4 APPLICATION 1 : COUPON REPLICATIONS

We consider the coupon replication model introduced in [20] and [19]. This application demonstrates that there are (exceptional) cases where an explicit expression can be obtained for the correction term V_i of the refined mean field model.

In such a system there are K different coupons and users arrive with coupon i at rate λ_i. Users are characterized by their current collection of coupons, and leave the system as soon as they collected the K coupons. Users regularly sample the population of users, either by sampling a user that currently holds the same number of coupons (i.e., the layered model) or by sampling any user (i.e., the flat model). If the sampled user holds at least one coupon that is not owned by the sampling user, the latter adds such a coupon to his collection.

4.1 Layered Model

In this subsection we focus on a single layer of the layered model (see [20, Section 3]). More specifically we discuss the dynamics of the first layer as the ODE of the other layers has exactly the same structure. Let $K > 2$ be the number of coupons and λ_i the rate of users arriving in layer 1 with coupon i. Let $x_i(t)$ be the rescaled number of users in the first layer at time t that hold coupon i. The transitions of the system are (for $i = 1, \ldots, K$):

$$\begin{aligned}&x \mapsto x + \frac{1}{N} e_i \text{ at rate } \beta_{e_i}(x) = \lambda_i, \\ &x \mapsto x - \frac{1}{N} e_i \text{ at rate } \beta_{-e_i}(x) = x_i \left(1 - \frac{x_i}{\sum_{k=1}^K x_k}\right).\end{aligned}$$

The first line corresponds to the arrival of a user with a coupon i and the second line to a user leaving to the next layer because he found a second coupon. The set of ODEs is therefore given by

$$\frac{d}{dt}x_i(t) = f_i(x) = \lambda_i - x_i(t)\left(1 - \frac{x_i(t)}{\sum_{k=1}^K x_k(t)}\right).$$

Let π be the unique fixed point of f. We have

$$f_i(x+\pi) = \lambda_i - (\pi_i + x_i)\left(1 - \frac{(\pi_i + x_i)}{\sum_{k=1}^K (\pi_k + x_k)}\right)$$

for $1 \le i \le K$.

Let $\Pi = \sum_{i=1}^K \pi_i$, then we find

$$A_{i,j} = \frac{\partial f_i}{\partial x_j}(\pi) = \begin{cases} -(\Pi - \pi_i)^2/\Pi^2 & i = j, \\ -\pi_i^2/\Pi^2 & i \neq j, \end{cases} \tag{12}$$

and

$$(B_j)_{k_1,k_2} = \frac{\partial f_j}{\partial x_{k_1}\partial x_{k_2}}(\pi) = \begin{cases} 2\pi_j^2/(\lambda\Pi^3) & j = k_1 = k_2, \\ 2(\Pi - \pi_j)^2/(\lambda\Pi^3) & j \neq k_1, k_2, \\ 2\pi_j(\Pi - \pi_j)/(\lambda\Pi^3) & j = k_1 \neq k_2, \\ 2\pi_j(\Pi - \pi_j)/(\lambda\Pi^3) & j = k_2 \neq k_1. \end{cases}$$

Symmetric case: From now on we focus on the symmetric case $\lambda_i = \lambda$ for all $i \in \{1, \dots, K\}$ (which was proven to minimize the mean sojourn times in [20]). In this case $\pi_i = \lambda K/(K-1)$, meaning the number of users in the system with coupon i is estimated by the classic mean field limit as $\lambda NK/(K-1)$. Due to (12), A is given by

$$A = -\frac{1}{K^2}\left(K(K-2)I + ee^T\right),$$

where I is the identity matrix and e a column vector of ones. As A is the sum of a diagonal matrix and a rank one matrix, the Sherman-Morrison-Woodbury formula implies

$$(A^{-1})_{i,j} = \frac{1}{(K-1)(K-2)} - \frac{\mathbf{1}_{\{i=j\}}K}{K-2},$$

where $\mathbf{1}_{\{Z\}} = 1$ if Z is true and $\mathbf{1}_{\{Z\}} = 0$ otherwise. The matrix Q is a diagonal matrix with all its entries equal to -2λ. As A is symmetric the unique solution W of the Lyapunov equation $AX + XA^T + Q = 0$ is given by λA^{-1}. Some basic algebra shows that

$$\sum_{k_1,k_2} (B_j)_{k_1,k_2} W_{k_1,k_2} = -\frac{2(K-1)^2}{(K-2)K^2},$$

and

$$\frac{1}{2}\sum_j (A^{-1})_{i,j} \sum_{k_1,k_2} (B_j)_{k_1,k_2} W_{k_1,k_2} = \frac{(K-1)}{K(K-2)}.$$

This implies that the mean field correction term given by (3) has the following simple explicit form

$$V_i = \frac{(K-1)}{K(K-2)},$$

and the refined mean field model suggest that the mean number of users that hold coupon i in the symmetric system with $\lambda_i = \lambda N$ is approximately

$$\frac{\lambda NK}{K-1} + \frac{K-1}{K(K-2)}.$$

K	N	$\pi_i N$	$\pi_i N + V_i$	simulation
5	10	1.250	1.517	1.530
	20	2.500	2.767	2.772
	50	6.250	6.517	6.522
8	10	1.143	1.289	1.291
	20	2.286	2.432	2.435
	50	5.714	5.860	5.861

Table 2. Mean number of users with coupon i in the symmetric layered model with arrival rate $\lambda_i N = N/10$

In Table 2 we present a comparison of the classic and refined mean field result with simulation. For this simulation users select a user uniformly at random after an exponential amount of time with mean one and we allow users to select themselves. We observe that the refined model is very close to the result obtained by simulation. For the system where we demand that the selected user is another user, the rate of the transitions with $\boldsymbol{\ell} = -e_i$ becomes dependent on N and one finds

$$\lim_{N\to\infty} N(\beta^{(N)}_{-e_i}(\pi) - \beta_{-e_i}(\pi)) = -\frac{\pi_i \sum_{k\neq i} \pi_k}{\left(\sum_{k=1}^K \pi_k\right)^2},$$

which equals $-(K-1)/K^2$ in the symmetric case. As the row sum of $-A^{-1}$ equals $K/(K-1)$ we get the following refined mean field approximation for the number of users in the system holding coupon i in case a user always selects *another* user

$$\frac{\lambda NK}{K-1} + \frac{K-1}{K(K-2)} - \frac{1}{K} = \frac{\lambda NK}{K-1} + \frac{1}{K(K-2)},$$

meaning the correction term with respect to the classic mean field reduces by a factor $K-1$.

4.2 Flat Model

We now consider the flat model considered in [19, Section 3.1] with $m = 1$, which is a reformulation of the model presented in [20, Section 4]. Let $K > 2$ be the number of coupons that a user needs to collect and λ the rate of users arriving with a single coupon. Let $y_i(t)$ be the fraction of the total number of users in the system at time t that hold i coupons. Define

$$q_i(y) = \sum_{j=1}^{K-1} p_{i,j} \frac{y_j}{\sum_{u=1}^{K-1} y_u},$$

where $p_{i,j} = 1$ if $j > i$ and $p_{i,j} = 1 - C^i_j/C^K_j$, for $j \leq i$, with C^b_a the number of ways to select a different items from a set of b different items.

The density dependent Markov model contains the following transitions: $\beta_{e_1}(y) = \lambda$, $\beta_{e_{i+1}-e_i}(y) = y_i q_i(y)$, for $i = 1, \ldots, K-1$, and $\beta_{-e_{K-1}}(y) = y_{K-1} q_{K-1}(y)$. Hence, the set of ODEs is given by

$$\frac{d}{dt} y_i(t) = y_{i-1}(t) q_{i-1}(y(t)) - y_i(t) q_i(y(t)),$$

for $i = 2, \ldots, K-1$ and $\frac{d}{dt} y_1(t) = \lambda - y_1(t) q_1(y(t))$. We compute a fixed point[2] of this set of ODEs using [20, Proposition 3] and note that for any fixed point π we have $\lambda = \pi_i q_i(\pi)$, for

[2]Global attraction of this fixed point is still an open issue.

N	$\pi_1 N$	$\pi_1 N + V_1$	simulation	$\pi_9 N$	$\pi_9 N + V_9$	simulation
10	1.009	1.018	1.019	1.839	1.752	1.752
20	2.019	2.028	2.024	3.678	3.591	3.588
50	5.047	5.056	5.062	9.194	9.107	9.100

Table 3. Mean number of users with $i = 1$ and $i = 9$ coupons in the flat model with arrival rate $\lambda = N/10$ and $K = 10$

$i = 1, \dots, K-1$. Using this equality and denoting $\sum_i \pi_i = \Pi$ one finds (for any fixed point)

$$A_{i,j} = \frac{\mathbf{1}_{\{i>1\}} p_{i-1,j} \pi_{i-1}}{\Pi} - \frac{p_{i,j}\pi_i}{\Pi} + \mathbf{1}_{\{i=1\}} \frac{\lambda}{\Pi} - \mathbf{1}_{\{i=j\}} \frac{\lambda}{\pi_i} + \mathbf{1}_{\{i-1=j\}} \frac{\lambda}{\pi_{i-1}},$$

and

$$\begin{aligned}(B_j)_{k,s} =& (p_{j,k} + p_{j,s}) \frac{\pi_j}{\Pi^2} - (p_{j-1,k} + p_{j-1,s}) \frac{\pi_{j-1}}{\Pi^2} \mathbf{1}_{\{j>1\}} - \mathbf{1}_{\{j=1\}} \frac{2\lambda}{\Pi^2} + \frac{\mathbf{1}_{\{k=j\}}}{\Pi} \left(\frac{\lambda}{\pi_j} - p_{j,s} \right) \\ &+ \frac{\mathbf{1}_{\{s=j\}}}{\Pi} \left(\frac{\lambda}{\pi_j} - p_{j,k} \right) - \frac{\mathbf{1}_{\{k=j-1\}}}{\Pi} \left(\frac{\lambda}{\pi_{j-1}} - p_{j-1,s} \right) - \frac{\mathbf{1}_{\{s=j-1\}}}{\Pi} \left(\frac{\lambda}{\pi_{j-1}} - p_{j-1,k} \right).\end{aligned}$$

Note that for $j > i > 1$ we have $A_{i,j} = (\pi_{i-1} - \pi_i)/\Pi$ and for $k, s > j > 1$ we have $(B_j)_{k,s} = 2(\pi_j - \pi_{j-1})/\Pi^2$. Further, the entries of A are independent of λ and the ones of B are inversely proportional to λ. As Q is a tridiagonal matrix with $Q_{n,n} = -2\lambda$ and $Q_{n,n+1} = Q_{n,n-1} = \lambda$, we note that the entries of W are proportional to λ and therefore the entries V_i are independent of λ as in the layered model.

Table 3 illustrates the accuracy of the refined mean field model in case $K = 10$ for $i = 1$ and $i = K-1 = 9$. Similar to the layered model we observe that the refined model is much more accurate that the classic mean field limit especially for small N.

5 APPLICATION 2 : THE SUPERMARKET MODEL

In this section, we consider the supermarket model introduced in [23, 29] where each job is allocated a server according to the "power-of-two-choices" algorithm. This model – or variants of this model – have been heavily studied in the literature and often serve as a basic example of a mean field interaction system. In this section, we show how our theory can be used to refine the analysis of this model. We also show that our refined approximation is accurate enough to differentiate between choices with replacement and without replacement.

5.1 The supermarket model

We consider a system composed of N identical servers. Jobs arrive at the system according to a Poisson process of rate ρN. The service time of each job is exponentially distributed of mean 1. For each incoming job, two servers are picked at random. The jobs is allocated to the server with the smallest queue size.

Let $X_i^{(N)}(t)$ denote the fraction of servers with queue size at least i at time t. As the role of each server is symmetrical, $X^{(N)}(t)$ is a Markov chain whose transitions are as follows. A departure from a server with $i \geq 1$ jobs modifies X into $X - N^{-1}\mathbf{e}_i$ and occurs at rate $N(X_i - X_{i+1})$. An arrival at a queue with i jobs modifies X into $X + N^{-1}\mathbf{e}_i$. The least loaded among the two has $i-1$ jobs if both servers have more than $i-1$ jobs but not both have i jobs. This probability depends on whether the two servers are picked with replacement or without replacement:

- If one picks two servers with replacement, this occurs with probability $X_{i-1}^2 - X_i^2$.

- If one picks two servers without replacement, this occurs with probability $X_{i-1}(NX_{i-1} - 1)/(N-1) - X_i(NX_i - 1)/(N-1)$

This leads to two (slightly) different models. For the classical two-choice model (with replacement), the transitions are, for $i \geq 1$:

$$\begin{array}{lll} X \to X - \frac{1}{N}\mathbf{e}_i & \text{at rate} & N(X_i - X_{i+1}) \\ X \to X + \frac{1}{N}\mathbf{e}_i & \text{at rate} & N\rho(X_{i-1}^2 - X_i^2) \end{array} \tag{13}$$

The transition of the model without replacement are:

$$\begin{array}{lll} X \to X - \frac{1}{N}\mathbf{e}_i & \text{at rate} & N(X_i - X_{i+1}) \\ X \to X + \frac{1}{N}\mathbf{e}_i & \text{at rate} & N\rho(X_{i-1}\frac{NX_{i-1}-1}{N-1} - X_i\frac{NX_i-1}{N-1}) \end{array} \tag{14}$$

The drift of the classical model (Equation (13)) does not depend on N and its ith coordinate is:

$$f_i(x) = \rho(x_{i-1}^2 - x_i^2) + (x_{i+1} - x_i). \tag{15}$$

This drift is also the limit of the model without replacement (Equation (14)) which shows that the mean field approximation of both models coincides as N goes to infinity. This explains why in the literature, there is no distinction made between the two models : the classical convergence result cannot differentiate the two models. In fact, the error between Equation (13) and Equation (14) is of order $O(1/N)$, which is of the same order as our refined term of Theorem 3.1. We will explore this in more detail in Section 5.3 to show that our results allows one to study the impact of choosing with or without replacement.

5.2 Refined Model for the Classical Two-Choice

In this section, we first study the refined model for the classical two-choice model (*i.e.*, with replacement Equation (13)). Its drift is given by Equation (15). It has a unique fixed point $\pi = (1, \pi_1, \pi_2, \dots)$ given by $\pi_i = \rho^{2^i-1}$.

5.2.1 Computations of A, B and W. To apply Theorem 3.1, the first step is to compute the Jacobian and Hessian of f. To compute the Jacobian and Hessian of f, let us denote by $\epsilon_i = x_i - \pi_i$. Developing $f(\pi + \epsilon)$, when $\epsilon \to 0$, we obtain for $i \geq 1$: (note: some terms cancel because $f(\pi) = 0$)

$$\begin{aligned} f_i(\pi + \varepsilon) &= 2\rho(\pi_{i-1}\epsilon_{i-1} - \pi_i\epsilon_i) + \epsilon_{i+1} - \epsilon_i + \rho(\epsilon_{i-1}^2 - \epsilon_i^2) \\ &= 2\pi_{i-1}\epsilon_{i-1} - (2\rho\pi_i + 1)\epsilon_i + \epsilon_{i+1} + \rho(\epsilon_{i-1}^2 - \epsilon_i^2). \end{aligned}$$

The Jacobian and Hessian in π are (for $i, j, k_1, k_2 \geq 1$):

$$A_{i,j} = \frac{\partial f_i}{\partial x_j}(0) = \begin{cases} 2\rho\pi_{i-1} = 2\rho^{2^{i-1}} & \text{if } j = i-1 \\ -2\rho\pi_i - 1 = -2\rho^{2^i} - 1 & \text{if } j = i \\ 1 & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

$$(B_j)_{k_1,k_2} = \frac{\partial^2 f_j}{\partial x_{k_1}\partial x_{k_2}}(0) = \begin{cases} -2\rho & \text{if } k_1 = k_2 = j \\ 2\rho & \text{if } k_1 = k_2 = j-1 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

The matrix Q of Equation (1) is: $Q = \sum_{\boldsymbol{\ell} \in \mathcal{L}} Q_{\boldsymbol{\ell}}$ with $(Q_{\boldsymbol{\ell}})_{n,m} = -\ell_n\ell_m\beta_{\boldsymbol{\ell}}(\pi)$. For the 2-choice model, this matrix is a diagonal matrix whose (i,i)-entry is

$$Q_{i,i} = -(\beta_{e_i}(\pi) + \beta_{-e_i}(\pi)) = -2(\rho^{2^i-1} - \rho^{2^{i+1}-1})$$

Let W be the (unique) solution of the Lyapunov equation $AX + XA^T + Q = 0$. The constant V_i of Theorem 3.1 is (for $i \geq 1$):

$$\begin{aligned} V_i &= \frac{1}{2}\sum_j (A^{-1})_{i,j} \sum_{k_1,k_2} (B_j)_{k_1,k_1} W_{k_1,k_2}, \\ &= \frac{1}{2}\sum_j A^{-1}_{i,j} 2\rho(-W_{j,j} + W_{j-1,j-1}) \\ &= \rho \sum_j (A^{-1}_{i,j} - A^{-1}_{i,j+1}) W_{j,j} \end{aligned} \tag{18}$$

It is not hard to verify that the general term of $A^{-1}_{i,j} - A^{-1}_{i,j+1}$ is

$$A^{-1}_{i,j} - A^{-1}_{i,j+1} = \begin{cases} 2^{i-j-1}\rho^{2^i - 2^{j+1}} & \text{if } j < i \\ 0 & \text{otherwise.} \end{cases}$$

It is unclear whether the matrix W can be computed in closed form.

5.2.2 Average queue length. In Table 4, we display the average queue length, computed by simulation or computed by using the mean field approximation plus $\sum_i V_i/N$, where V_i is the constant given by Equation (18). Again, we observe that the refined mean field approximation greatly improves the quality of the approximation. Even for $N = 10$, the refined model has a relative error below 5%. Moreover, as the error of refined-mean field decreases as $O(1/N^2)$, the refined model quickly becomes very accurate.

	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = \infty$
Simulation ($\rho = 0.7$)	1.2194	1.1735	1.1584	1.1471	1.1384	–
Refined mean field	1.2150	1.1726	1.1584	1.1471	1.1386	1.1301
Simulation ($\rho = 0.9$)	2.8040	2.5665	2.4907	2.4344	2.3931	–
Refined mean field	2.7513	2.5520	2.4855	2.4324	2.3925	2.3527
Simulation ($\rho = 0.95$)	4.2952	3.7160	3.5348	3.4002	3.3047	–
Refined mean field	4.1017	3.6578	3.5098	3.3915	3.3027	3.2139

Table 4. Two-choice model : Average queue length for various values of ρ and N. We compare simulation with the refined mean field approximation

We confirm the observation made in [9] that the error made by the classical mean field approximation seems to grow as $1/(1-\rho)$ as the load ρ approaches 1. In [9], it was conjectured that the constant $\sum_i V_i/N$ was close (but not equal to) $\rho^2/(2-2\rho)$. Our numerical evaluation confirms this tendency.

The results in Table 4 also suggest that the difference between the mean field approximation and the values obtained by simulation again grows quickly as ρ approaches 1. By Theorem 3.2, it is possible to develop a second order approximation of the form $\pi + V/N + V^{(2)}/N^2$. By using a numerical fitting of the parameter, we propose to approximate the average queue length by the quantity $a(N,\rho)$, defined by

$$a(N,\rho) = \sum_{i=1}^{\infty} \pi_i(\rho) + \frac{\rho^2}{2N(1-\rho)} + \frac{1}{20N^2(1-\rho)^2}. \tag{19}$$

The above formula suggests that error of the refined approximation grows as $1/(1-\rho)^2$ as ρ approaches 1.

In Table 5, we provide some numerical results showing the accuracy of this second order approximation against simulation result. This new approximation appears to be very accurate for all tested values.

	$N=10$	$N=20$	$N=30$	$N=50$	$N=100$
Simulation ($\rho=0.7$)	1.2194	1.1735	1.1584	1.1471	1.1384
$a(N,\rho)$	1.2173	1.1723	1.1580	1.1467	1.1383
Simulation ($\rho=0.9$)	2.8040	2.5665	2.4907	2.4344	2.3931
$a(N,\rho)$	2.8077	2.5677	2.4932	2.4357	2.3937
Simulation ($\rho=0.95$)	4.2952	3.716	3.5348	3.4002	3.3047
$a(N,\rho)$	4.3164	3.7151	3.5369	3.4024	3.3061

Table 5. Two-choice model : Average queue length for various values of ρ and N. We compare simulation with a second order approximation defined by Equation (19).

5.2.3 Distribution of queue size. In Figure 1, we plot the marginal law of the number of jobs at a server. We compare the mean field approximation (dashed line), the simulation results and the refined mean field approximation, given by $\pi_i + V_i/N$, where V_i is given by Equation (18). We restrict our observation to $N = 50$ and $\rho = 0.95$, but other values are similar.

In Figure 1(a), we observe that the refined approximation predicts the shape of the distribution with great accuracy : for $N = 50$, the curves are almost indistinguishable. The situation of the tail is different. In Figure 1(b), we plot in log-scale the quantity $\mathbb{E}[X_i^{(N)}] = \mathbf{P}[\text{queue length} \geq i]$ as a function of i. We observe that while the refined mean field approximation greatly improves the accuracy for the average response time or the shape of the distribution, it does not accurately predict the tail behavior: it is an improvement compared to the mean field approximation, but remains quite inaccurate. This can be explained as follows. Theorem 3.1 guarantees that, uniformly in i,

$$\left|\pi_i + \frac{V_i}{N} - \mathbf{P}\left[\textit{queue length} \geq i\right]\right| = O\left(\frac{1}{N^2}\right).$$

However, the relative error $\left|\pi_i + V_i/N - \mathbf{P}[\textit{queue length} \geq i]\right| / \mathbf{P}[\textit{queue length} \geq i]$ is not guaranteed by this result to remain small as i grows because the quantity $\mathbf{P}[\textit{queue length} \geq i]$ converges quickly to 0 as i grows.

5.3 Impact of Choosing Without Replacement

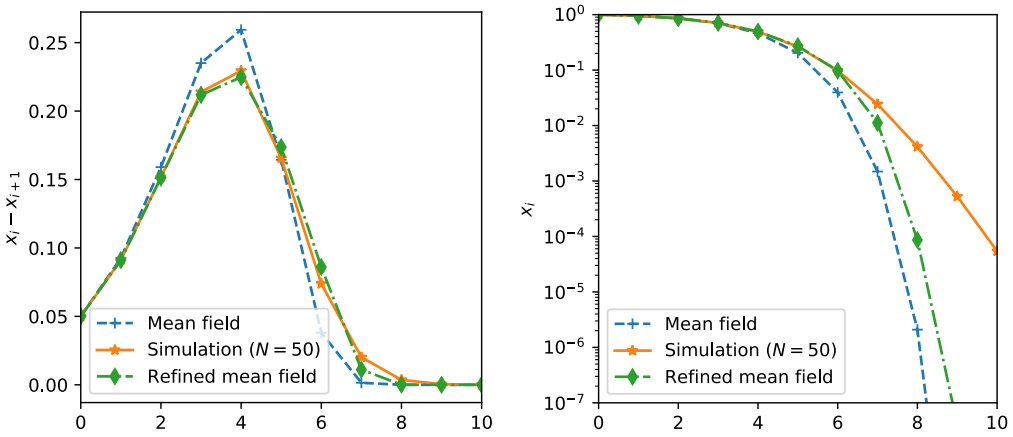
We now focus on the comparison of the two-choice model with replacement (Equation (13)) and without replacement (Equation (14)). The only difference between the two models is the rate of arrivals. Rewriting Equation (13) and Equation (14), the rate of a transition $X \mapsto X + e_i/N$ (for $i \geq 1$) is one of the following:

$$\beta_i(x) = \rho(x_{i-1}^2 - x_i^2) \qquad \text{(with replacement, Eq. (13))}$$

$$\beta_i^{(N)}(x) = \rho(x_{i-1} - x_i)(x_{i-1} + x_i - \frac{1}{N})\frac{N}{N-1} \qquad \text{(without replacement, Eq. (14))}$$

For each i, the function $\beta_i^{(N)}(x)$ converges to the function $\beta_i(x)$ as N goes to infinity. Moreover, we have

$$\lim_{N\to\infty} N(\beta_i^{(N)}(x) - \beta_i(x)) = \rho(x_{i-1} - x_i)(x_{i-1} + x_i - 1).$$



(a) queue-length distribution (b) tail behavior (log-scale).

Fig. 1. Two-choice model : Distribution of the queue length estimated by simulation, the mean field approximation and its refined version.

Theorem 3.1 shows that, for a coordinate i, the expected number of servers with a queue size i or more, $\mathbb{E}[X_i]$, can be expressed as

$$\mathbb{E}\left[X_i^{(N)}\right] = \pi_i + \frac{V_i}{N} + \frac{E_i}{N} + O\left(\frac{1}{N^2}\right), \tag{20}$$

where V_i is the refined constant of the classical two-choice (with replacement), defined in Equation (18) and E_i is defined by :

$$\begin{aligned} E_i &= -\sum_{j=1}^{\infty} A^{-1}_{i,j}\rho(\pi_{j-1}-\pi_j)(\pi_{j-1}+\pi_j-1) \\ &= \sum_{j=1}^{\infty} \rho(\pi_j^2-\pi_j)(A^{-1}_{i,j+1}-A^{-1}_{i,j}) \\ &= \sum_{j=1}^{i-1}(1-\pi_j)2^{i-j-1}\rho^{2^i-2^j} \end{aligned}$$

5.3.1 Average queue length. For the average queue length, the difference is the sum over all i of the above expression. This shows that for large N, the average queue length of the two-choice model without replacement will be E/N smaller than the one of the two-choice model with replacement, where

$$E = \sum_{i=1}^{\infty} E_i = \sum_{i=1}^{\infty}\sum_{j=1}^{i-1}(1-\pi_j)2^{i-j-1}\rho^{2^i-2^j} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty}(\rho^{2^{i+j}-2^j}-\rho^{2^{i+j}-1})2^{i-1}$$

As an illustration, we compare in Table 6 the refined approximation and values obtained by simulation for various values of ρ and N. We observe that in most cases, the error E/N is a good predictor of the impact of choosing with or without replacement.

5.3.2 Error of the mean field approximation v.s. *error of the model with/without replacement.* As indicated by Equation (20), the average queue length for the two-choice model without replacement

	$N = 10$	$N = 20$	$N = 50$
$\rho = 0.7$			
with replacement	1.217 (rmf=1.215)	1.173 (rmf=1.173)	1.147 (rmf=1.147)
without replacement	1.171 (rmf=1.169)	1.151 (rmf=1.150)	1.137 (rmf=1.138)
with-without	0.047 (rmf=0.046)	0.022 (rmf=0.023)	0.010 (rmf=0.009)
$\rho = 0.9$			
with replacement	2.055 (rmf=2.751)	2.574 (rmf=2.552)	2.431 (rmf=2.432)
without replacement	1.957 (rmf=2.630)	2.502 (rmf=2.491)	2.406 (rmf=2.408)
with-without	0.098 (rmf=0.121)	0.073 (rmf=0.061)	0.026 (rmf=0.024)
$\rho = 0.95$			
with replacement	4.198 (rmf=4.102)	3.699 (rmf=3.658)	3.406 (rmf=3.391)
without replacement	4.015 (rmf=3.923)	3.600 (rmf=3.568)	3.366 (rmf=3.356)
with-without	0.184 (rmf=0.179)	0.099 (rmf=0.089)	0.040 (rmf=0.036)

Table 6. Two-choice : comparison with and without replacement. The values in parenthesis (rmf=·) correspond to the refined mean field approximation, the others to the values obtained by simulation. The rows "with-without" are the difference between "with replacement" and "without replacement".

can be expressed as $\sum_i \pi_i + V/N + E/N + O(1/N^2)$ where $\sum_i \pi_i$ is the mean field approximation and $\sum_i \pi_i + V/N$ is the refined approximation for the classical model with replacement.

In Figure 2, we compare the constant V and E by plotting them as a function of ρ. We plot, in solid blue the refined constant V and in dashed orange the term E that corresponds to choosing with replacement and without replacement. We observe that the constant V is always larger than E. Moreover, when ρ is close to 1, the error due to with or without replacement grows, but grows much slower than the correction term V. Numerically, we observe that E seems to grow as $\log(1/(1-\rho))$ as ρ goes to one (similar to the average queue length) whereas V seems to grow as $1/(1-\rho)$.

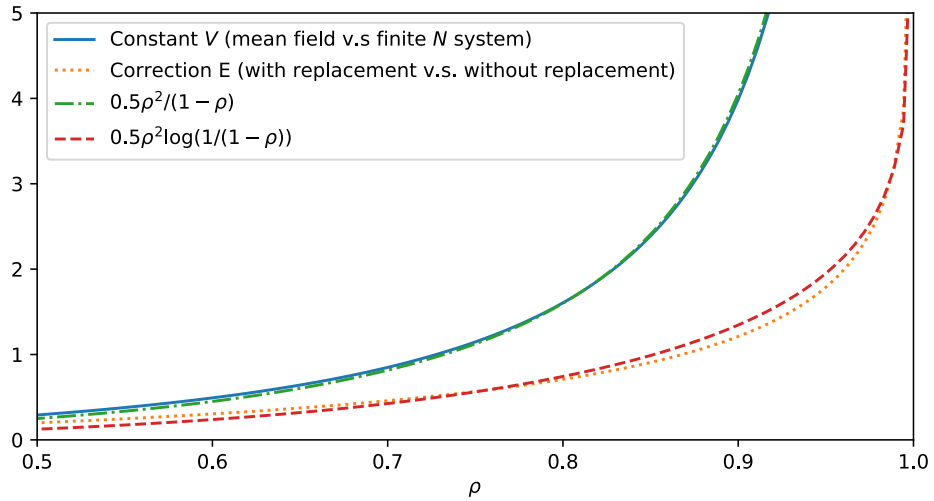


Fig. 2. Two-choice : Comparison of the constants V (error of the mean field) and E (difference between choosing with or without replacement).

6 APPLICATION 3 : PUSH-PULL STRATEGIES

In this section we illustrate the refined mean field approximation on a model used to study a class of pull and push strategies in a distributed network consisting of N servers. The main reason for considering this application is that it illustrates that the refined mean field model can drastically improve the accuracy of the classic mean field model where for $N = 10$ errors close to 100% reduce to less than 1% (see Table 7).

The push-pull model is as follows. Jobs arrive at each server according to a Poisson process. Each server has a single server with an infinite waiting room and serves jobs in FCFS order. Under a pull strategy idle servers sample a server uniformly at random at rate r and if the sampled server holds at least 2 jobs, a job is exchanged. In case of a push strategy the servers with pending jobs sample servers in order to find an idle server. As shown in [21], the performance of both the pull and push strategy can be captured using the mean field model considered in this section.

Let x_k be the fraction of the servers with k or more jobs. There are three types of transitions (arrivals, departures or job transfers):

$$\begin{aligned}\beta_{e_k}(x) &= \lambda(x_{k-1} - x_k),\\ \beta_{-e_k}(x) &= (x_k - x_{k+1}),\\ \beta_{e_1-e_k}(x) &= r(1-x_1)(x_k - x_{k+1}).\end{aligned}$$

One easily verifies that (as π is a fixed point)

$$\begin{aligned}f_1(x+\pi) &= -(1+\lambda+r\pi_2)x_1 + (1+r(1-\lambda)r)x_2 - rx_1x_2,\\ f_i(x+\pi) &= \lambda x_{i-1} - (\lambda+1+(1-\lambda)r)x_i + (1+(1-\lambda)r)x_{i+1} + (\pi_i - \pi_{i+1})rx_1 + rx_1(x_i - x_{i+1}),\end{aligned}$$

for $i > 1$.

Define $\rho = \lambda/(1+(1-\lambda)r)$, such that $\pi_i = \lambda\rho^{i-1}$ for $i \geq 1$, then

$$\frac{\rho}{\lambda}A_{i,j} = \begin{cases} -\rho - \rho^2 r - \rho/\lambda & i = j = 1,\\ \rho + (1-\rho)\rho^2 r & i = 2, j = 1,\\ (1-\rho)\rho^i r & i > 2, j = 1,\\ 1 & j = i+1,\\ -(1+\rho) & j = i > 1,\\ \rho & j = i-1 > 0,\\ 0 & \text{otherwise,}\end{cases}$$

and

$$(B_j)_{k_1,k_2} = \begin{cases} -r & j \geq 1, (k_1,k_2) = (1, j+1) \text{ or } (j+1, 1)\\ r & j > 1, (k_1,k_2) = (1,j) \text{ or } (j,1)\\ 0 & \text{otherwise.}\end{cases}$$

Note that the transposed matrix A^T is the subgenerator matrix of a phase-type distribution as its diagonal entries are negative, its off-diagonal entries are nonnegative and its row sums are all equal to zero, except for the first row which sums to minus one [17]. Hence, the eigenvalues of A all have negative real parts and the Lyapunov equation $AX + XA^T + Q = 0$ has a unique solution.

The sum formula (3) therefore simplifies to

$$V_i = -r\left(\sum_{j\geq 1}(A^{-1})_{i,j}W_{1,j+1} - \sum_{j>1}(A^{-1})_{i,j}W_{1,j}\right)$$
$$= -r\left(\sum_{j\geq 1}\left[(A^{-1})_{i,j} - (A^{-1})_{i,j+1}\right]W_{1,j+1}\right),$$

as W is symmetric. We now argue that

$$(A^{-1})_{i,j} - (A^{-1})_{i,j+1} = \begin{cases} \rho^{i-j}/\lambda & i > j, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

A^T is the subgenerator matrix of a phase-type distribution, as such entry (j,i) of $(-A^T)^{-1}$ is equal to the expected time spent in phase i given that we start in state j until absorption. When $i \leq j$ it is clear that the expected time spent in phase i remains the same whether we start from state j or $j+1$ (as any path from phase $j+1$ to phase i must visit phase j first). Therefore $(A^{-1})_{i,j} - (A^{-1})_{i,j+1} = 0$ for $i \leq j$. When $i > j$ we find that $(A^{-1})_{i,j} - (A^{-1})_{i,j+1} = ((-A^T)^{-1})_{j+1,i} - ((-A^T)^{-1})_{j,i}$ which is the expected time spent in phase i starting from phase $j+1$ under taboo of phase j. This expectation is given by $\rho^{i-(j+1)}/(\lambda/\rho)$ due to the birth-death structure [17].

Due to (21) we obtain

$$V_i = \frac{-r}{\lambda}\left(\sum_{j=2}^{i}\rho^{i-j+1}W_{1,j}\right),$$

with W the unique solution of the Lyapunov equation $AX + XA^T + Q = 0$ with $Q = Q_a + Q_d + Q_t$ and

$$(Q_a)_{n,m} = \begin{cases} -\lambda(\pi_{k-1} - \pi_k) & n = m = k \geq 1, \\ 0 & \text{otherwise,} \end{cases}$$
$$(Q_d)_{n,m} = \begin{cases} -(\pi_k - \pi_{k+1}) & n = m = k \geq 1, \\ 0 & \text{otherwise,} \end{cases}$$
$$(Q_t)_{n,m} = \begin{cases} -r(1-\lambda)\pi_2, & n = m = 1, \\ -r(1-\lambda)(\pi_k - \pi_{k+1}) & n = m = k \geq 2, \\ r(1-\lambda)(\pi_k - \pi_{k+1}) & k \geq 2, (n,m) = (1,k) \text{ or } (k,1), \\ 0 & \text{otherwise.} \end{cases}$$

A comparison between the mean queue length of the refined mean field and simulation is presented in Table 7 when $r = 1/(1-\lambda)$ for various N and λ values. We remark that for $\lambda = 0.95$, the mean field approximation has relative errors close to 90%, while the refined model has relative errors below 0.5%.

Figure 3 illustrates the accuracy of the refined mean field approximation on the queue length distribution when $\lambda = 0.90$. Simulation results are based on a run of length 10^8. As for the supermarket model (see Figure 1(b)), we observe again that although the refined approximation greatly improves the accuracy of the estimation of the average response time compared to the classical mean field approximation, it is less accurate for the tail behavior. The explanation is similar to that of §5.2.3: in Figure 3, we observe that the relative error grows with i because it is of the order of $O(1/N)/\mathbf{P}[\textit{queue length} \geq i]$, but the absolute error is bounded uniformly in i.

λ	N	10	20	50	100	∞
0.80	Simulation	1.5569	1.4438	1.3761	1.3545	–
	Refined mean field	1.5473	1.4403	1.3761	1.3547	1.3333
0.90	Simulation	2.3043	1.9700	1.7681	1.7023	–
	Refined mean field	2.2945	1.9654	1.7680	1.7022	1.6364
0.95	Simulation	3.4288	2.6151	2.1330	1.9720	–
	Refined mean field	3.4369	2.6232	2.1350	1.9723	1.8095

Table 7. Mean queue length under pull/push with $r = 1/(1-\lambda)$: simulation vs refined mean field approximation

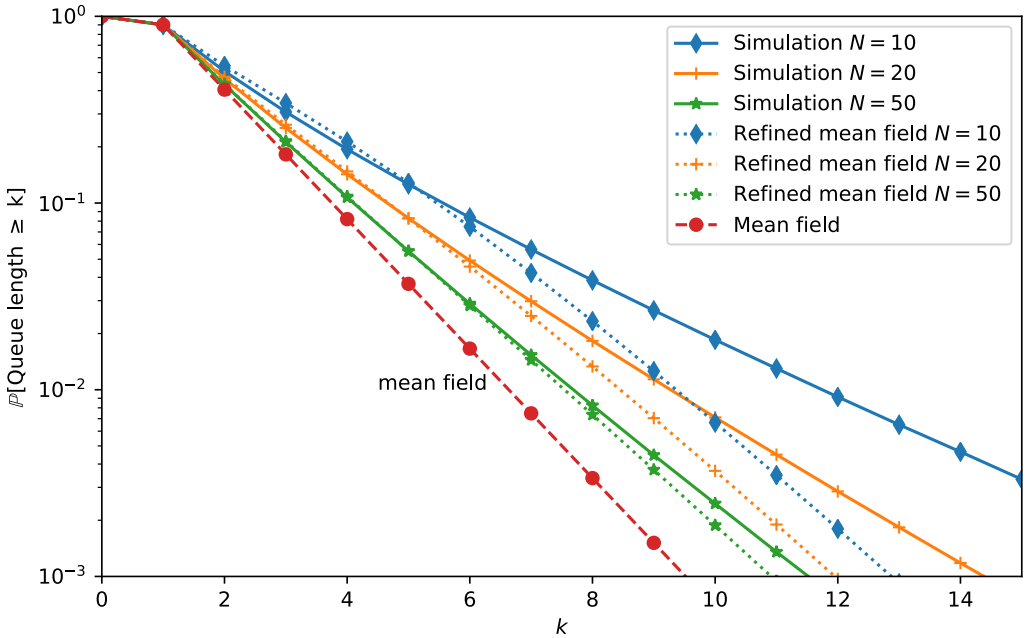


Fig. 3. Queue length distribution under pull/push with $r = 1/(1-\lambda)$: simulation vs refined mf. approximation.

7 APPLICATION 4 : BIN PACKING AND BLOCKING PROBABILITIES

In this subsection we consider the bin packing model of [30], where the total job arrival rate is λN and each server is a loss queue with capacity K. Arriving jobs pick 2 servers at random and join the queue with the largest spare capacity. If both selected queues have no spare capacity the incoming job is blocked and considered lost. While the focus was mainly on the mean response time in case of the previous applications, we now look at the accuracy of the refined mean field model when considering small blocking probabilities.

There are two types of transitions (arrivals and departures):

$$\beta_{e_k}(x) = \lambda(x_{k-1}^2 - x_k^2),$$
$$\beta_{-e_k}(x) = k(x_k - x_{k+1}),$$

for $k = 1, \ldots, K$, where $x_0 = 1$ and $x_{K+1} = 0$.

λ	K	N	10	50	100	∞
9	10	Simulation	7.512 E-2	5.997 E-2	5.777 E-2	–
		Refined mean field	7.753 E-2	6.002 E-2	5.783 E-2	5.565 E-2
		Relative error (refined mf)	-0.031	-0.0008	-0.001	–
		Relative error (mean field)	0.35	0.078	0.038	–
7	10	Simulation	6.584 E-3	8.382 E-4	3.997 E-4	–
		Refined mean field	2.450 E-3	5.589 E-4	3.226 E-4	8.632 E-5
		Relative error (refined mf)	1.7	0.5	0.23	–
		Relative error (mean field)	75	8.7	3.6	–
36	40	Simulation	8.501 E-3	8.207 E-4	2.162 E-4	–
		Refined mean field	1.865 E-4	3.778 E-5	1.918 E-5	5.893 E-7
		Relative error (refined mf)	45	21	10	–
		Relative error (mean field)	14420	1390	366	–

Table 8. Blocking probability: simulation (length: $\lambda 10^8$ arrivals) vs refined mean field approximation. The relative error are computed as (Value obtained by simulation)/(value of the [refined] mean field).

It is shown in [30] that the corresponding system of ODE has a unique fixed point π. Moreover, around this point π we have

$$\begin{aligned} f_1(x+\pi) &= -\lambda x_1^2 - 2\lambda x_1\pi_1 - (x_1 - x_2), \\ f_i(x+\pi) &= \lambda(x_{i-1}^2 - x_i^2) + 2\lambda(x_{i-1}\pi_{i-1} - x_i\pi_i) - i(x_i - x_{i+1}), \\ f_K(x+\pi) &= \lambda(x_{K-1}^2 - x_K^2) + 2\lambda(x_{K-1}\pi_{K-1} - x_K\pi_K) - Kx_K. \end{aligned}$$

for $1 < i < K$. Hence,

$$A = \begin{bmatrix} -2\lambda\pi_1 - 1 & 1 & & & \\ 2\lambda\pi_1 & -2\lambda\pi_2 - 2 & 2 & & \\ & 2\lambda\pi_2 & \ddots & \ddots & \\ & & \ddots & \ddots & K-1 \\ & & & 2\lambda\pi_{K-1} & -2\lambda\pi_K - K \end{bmatrix},$$

and

$$(B_j)_{k_1,k_2} = \begin{cases} -2\lambda & 1 \le j \le K, (k_1,k_2) = (j,j) \\ 2\lambda & 1 < j \le K, (k_1,k_2) = (j-1, j-1) \\ 0 & \text{otherwise} \end{cases}$$

The sum formula (3) can therefore be written as

$$V_i = -\lambda\left(\sum_{j=1}^{K-1}\left[(A^{-1})_{i,j} - (A^{-1})_{i,j+1}\right]W_{j,j} + (A^{-1})_{i,K}W_{K,K}\right),$$

with W the unique solution to $AX + XA^T + Q = 0$ where Q is a diagonal matrix with

$$Q_{i,i} = -\lambda(\pi_{i-1}^2 - \pi_i^2) - i(\pi_i - \pi_{i+1}),$$

for $i = 1, \ldots, K$, with $\pi_0 = 1$ and $\pi_{K+1} = 0$.

Table 8 illustrates the accuracy of the refined mean field model for various parameter settings for λ and K for the blocking probability. While the blocking probability of the classic mean field approximation can be several orders of magnitude below the simulation result, the refined mean

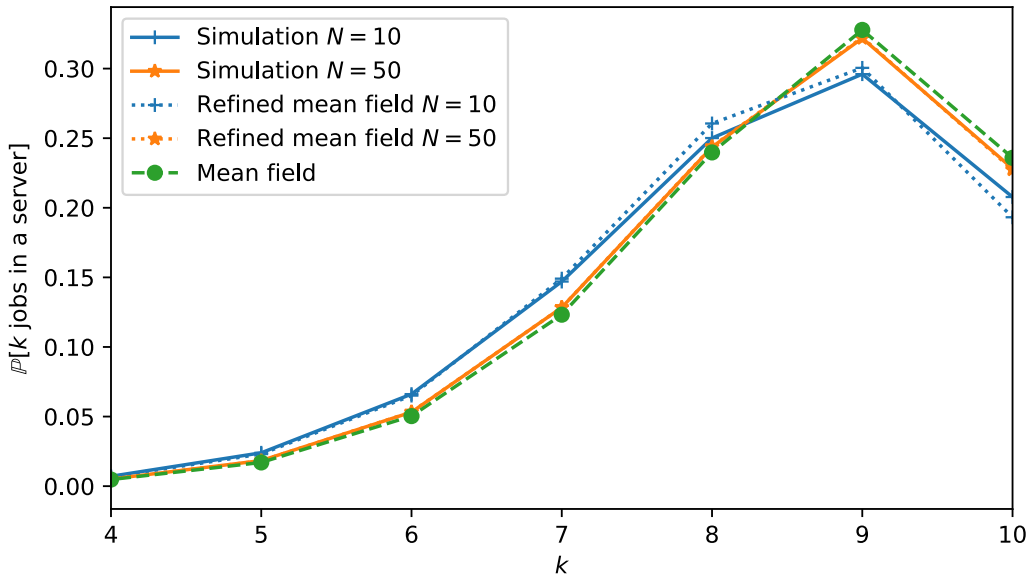


Fig. 4. Bin packing: Server occupancy distribution or $\lambda = 9$ and $K = 10$: simulation vs refined mf. approximation.

field model is significantly closer. Note that, as for the supermarket and the pull/push model, the refined mean field is an improvement compared to the classical mean field but does not fully predict the tail behavior: this explains why the relative error is larger when $K = 40$ than when $K = 10$. The blocking probability p_{blocked} of the refined mean field model was computed by noting that the rate at which jobs are processed should match the rate at which jobs are accepted into the system, *i.e.,*

$$\lambda(1 - p_{\text{blocked}}) = \sum_{i=1}^{K} (\pi_i + V_i/N).$$

For the classic mean field approximation this is equivalent to stating $p_{\text{blocked}} = \pi_K^2$ as the queue lengths are asymptotically independent, but this is not the case for finite N. This is further illustrated by Figure 4, which compares the server occupancy distribution of the refined mean field approximation with simulation for $\lambda = 9$ and $K = 10$. It shows that while the probability of having K jobs tends to decrease as we decrease N, the blocking probability nevertheless increases (see Table 8) due to the correlation in the joint queue lengths.

8 CONCLUSION

In this paper, we proposed a refined mean field approximation that is shown to be significantly more accurate than the classic mean field limit for systems composed of a limited number of objects. This refinement exists in approximating $E^{(N)}[h(X^{(N)})]$ by $h(\pi) + V_h/N$, where π is the fixed point (i.e., $h(\pi)$ is the classic mean field limit) and V_h is the limit of $N(E^{(N)}[h(X^{(N)})] - h(\pi))$ as N tends to infinity. We showed that the constant V_h can be expressed via the Jacobian and Hessian of the drift f in π and the solution of a single Lyapunov equation, which can be evaluated in less than a second for models with as many as 500 dimensions.

One of the key assumptions to establish our main result is that the drift f is twice-differentiable. For systems with a discontinuous drift (e.g., [27]) the decay of $E^{(N)}[h(X^{(N)})] - h(\pi)$ may be slower than $1/N$ if the fixed point is located at a point in which the drift is discontinuous. When the drift is up to k times differentiable, the refined approximation can in principle be further improved to obtain a $o(1/N^k)$ error. However, at this stage it is not clear whether an efficient computational method can be devised to compute the constants of the $1/N^i$ terms for $i = 2, \ldots, k$.

APPENDIX

A DISCRETE-TIME MODEL : A 802.11 EXAMPLE

In this section, we show how we can adapt our results to the model of the 802.11 MAC protocol presented in [6]. This models falls into the class of mean field interacting models of [2] and the purpose of this section is to show that one can easily apply our results to such models.

The authors of [6] presents a discrete-time model of the 802.11 MAC protocol. There are N transmitters and each transmitter can be in a state $\{0, \ldots, d-1\}$. At each time slot, a transmitter that is in a state k attempts to transmit with probability q_k/N. If it is the only one transmitting, then the transmission is successful and its state becomes 0 at the next time slot. If there are two or more nodes transmitting, then the state becomes $(k+1) \mod d$.

Let $X_i^{(N)}(t)$ denote the fraction of transmitters in state i at time t. If n is a node in state k, then the probability that at least one other node transmits at the same time is $\gamma_k(X^{(N)}(t))$, where:

$$\gamma_k^{(N)}(x) = 1 - (1 - \frac{q_k}{N})^{-1} \prod_{s=0}^{d-1} (1 - \frac{q_s}{N})^{Nx_s},$$

where Nx_s is the number of nodes in state s. This model falls into the class of models of [6] with an intensity $\varepsilon(N) = 1/N$.

By considering that the time steps last for a random time that is exponentially distributed with mean $1/N$, we obtain the following Markov process with transitions given by (for $k \in \{0, \ldots, d-1\}$)

$$\begin{aligned}
x &\mapsto x + \frac{1}{N}(\mathbf{e}_0 - \mathbf{e}_k) && \text{at rate } Nq_k x_k (1 - \gamma_k^N(x)),\\
x &\mapsto x + \frac{1}{N}(\mathbf{e}_{k+1} - \mathbf{e}_k) && \text{at rate } Nq_k x_k \gamma_k^N(x),\\
x &\mapsto x + \frac{1}{N}(\mathbf{e}_0 - \mathbf{e}_{d-1}) && \text{at rate } Nq_{d-1} x_{d-1}.
\end{aligned}$$

The first line corresponds to a successful transmission, the second to a collision. The last one is because any attempted transmission from state $d-1$ goes back to state 0 (successful or not). Note that this model has also some transitions that affect $k \geq 2$ objects but their rate is $O(N^{1-k})$ and they are therefore negligible.

Note that, as N goes to infinity, we have

$$\lim_{N\to\infty} \gamma_k^{(N)}(x) = \gamma(x) = 1 - e^{-\sum_{k=0}^{d-1} q_k x_k},$$

which shows that model falls into our density-dependent population process introduced in Section 2.1.

This gives the following drift equations, where $K = d - 1$:

$$\begin{aligned} f_0(x) &= \sum_{k=1}^{K-1} q_k x_k (1-\gamma(x)) - q_0 x_0 \gamma(x) + q_K x_K \\ &= \sum_{k=0}^{K} q_k x_k (1-\gamma(x)) - q_0 x_0 + q_K x_K \gamma(x), \\ f_k(x) &= \gamma(x) q_{k-1} x_{k-1} - q_k x_k, \qquad \text{for } k \geq 1. \end{aligned}$$

which is the ODE presented in [6, Equations (5-6)]. Its fixed point is given by the famous Bianchi's formula:

$$x_k = \frac{\gamma^k}{q_k} x_0 = \frac{\gamma^k}{q_k \sum_{j=0}^{K} \gamma^j / q_j},$$

where $\gamma = 1 - e^{-\sum_k q_k x_k}$.

To determine the refined mean field approximation, the first and the second derivative of the drift can be easily obtained using a symbolic computation toolbox like `SymPy`. Also, we note that this model has a correction term C due to the limit $\lim_{N\to\infty} \gamma_k^{(N)} = \gamma$.

B HIGHER ORDER TERMS

In this section, we sketch how the constants $D^{(2)} \dots D^{(k)}$ can be computed. By using Lemma 3.5, the proof of Lemma 3.4 can be adapted to show that, for a function h that is $k+1$ times differentiable, we have:

$$(L^{(N)}h)(x) = (\Lambda h)(x) + \frac{1}{2N} \sum_{\boldsymbol{\ell} \in \mathcal{L}} \beta_{\boldsymbol{\ell}}(x) D^2 h(x) \cdot (\boldsymbol{\ell}, \boldsymbol{\ell}) + \dots + \frac{1}{k! N^{k-1}} \sum_{\boldsymbol{\ell} \in \mathcal{L}} \beta_{\boldsymbol{\ell}}(x) D^k h(x) \cdot (\boldsymbol{\ell}, \dots, \boldsymbol{\ell})$$

As for Equation (11), by using the chain rule, the expression $\int_0^\infty D^k \Phi_t(\pi) dt$ can be expressed as an integral of functions that involve up to the kth derivatives of the drift f evaluated in π as well as some factors of the form

$$\int_0^\infty \exp(At)_{i_1,j_1} \exp(At)_{i_2,j_2} \dots \exp(At)_{i_k,j_k} dt. \tag{22}$$

B.1 Is a numerical evaluation possible?

To evaluate the expression $\int_0^\infty D^k \Phi_t(\pi) dt$, one suffers two drawbacks. First, the complexity of this expression grows quickly with k, and second it requires to generalize the Lyapunov equation (1) in order to evaluate numerically the integral (22).

In what follows, we show how to evaluate numerically this integral by using the Kronecker product $\oplus$. Let M be the $(d^k) \times (d^k)$ matrix defined by :

$$M = A \oplus A \oplus \dots \oplus A. \qquad (k \text{ times})$$

In terms of indices, for $k = 3$, this means that

$$M_{i_1 i_2 i_3, j_1 j_2 j_3} = A_{i_1 j_1} \delta_{i_2 j_2} \delta_{i_3 j_3} + \delta_{i_1 j_1} A_{i_2 j_2} \delta_{i_3 j_3} + \delta_{i_1 j_1} \delta_{i_2 j_2} A_{i_3 j_3}.$$

We have:

$$\exp(Mt) = \exp(At) \otimes \exp(At) \otimes \dots \otimes \exp(At). \qquad (k \text{ times})$$

Again, in terms of indices, for $k = 3$, this means that:

$$\exp(Mt)_{i_1 i_2 i_3, j_1 j_2 j_3} = \exp(At)_{i_1,j_1} \exp(At)_{i_2,j_2} \exp(At)_{i_3,j_3}.$$

As A is Hurwitz, so is M. This implies that $\int_0^\infty \exp(Mt)dt = -M^{-1}$. In particular, this shows that

$$\int_0^\infty \exp(At)_{i_1,j_1} \exp(At)_{i_2,j_2} \exp(At)_{i_3,j_3} dt = -(M^{-1})_{i_1 i_2 i_3, j_1 j_2 j_3}.$$

As a result, the terms of Equation (22) can be computed by inverting the $d^k \times d^k$-matrix M which can be done in $O(d^{3k})$. Note that for $k = 2$, this method is less efficient than solving the Lyapunov equation (which can be done in $O(d^3)$).

ACKNOWLEDGMENT

This work is partially supported by the EU project QUANTICOL, 600708.

REFERENCES

[1] R. H. Bartels and G. W. Stewart. 1972. Solution of the Matrix Equation AX + XB = C [F4]. *Commun. ACM* 15, 9 (Sept. 1972), 820–826. https://doi.org/10.1145/361573.361582
[2] Michel Benaim and Jean-Yves Le Boudec. 2008. A class of mean field interaction models for computer and communication systems. *Performance Evaluation* 65, 11 (2008), 823–838.
[3] Luca Bortolussi, Jane Hillston, Diego Latella, and Mieke Massink. 2013. Continuous approximation of collective system behaviour: A tutorial. *Performance Evaluation* 70, 5 (2013), 317–349.
[4] Anton Braverman and Jim Dai. 2017. Stein's method for steady-state diffusion approximations of $M/Ph/n + M$ systems. *The Annals of Applied Probability* 27, 1 (2017), 550–581.
[5] Anton Braverman, JG Dai, and Jiekun Feng. 2017. Stein's method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. *Stochastic Systems* 6, 2 (2017), 301–366.
[6] Jeong-woo Cho, Jean-Yves Le Boudec, and Yuming Jiang. 2012. On the asymptotic validity of the decoupling assumption for analyzing 802.11 MAC protocol. *IEEE Transactions on Information Theory* 58, 11 (2012), 6879–6893.
[7] Richard Datko. 1972. Uniform asymptotic stability of evolutionary processes in a Banach space. *SIAM Journal on Mathematical Analysis* 3, 3 (1972), 428–445.
[8] Jaap Eldering. 2013. *Normally Hyperbolic Invariant Manifolds—the Noncompact Case (Atlantis Series in Dynamical Systems vol 2)*. Springer, Berlin.
[9] Nicolas Gast. 2017. Expected values estimated via mean-field approximation are 1/N-accurate. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1, 1 (2017), 17.
[10] Nicolas Gast and Gaujal Bruno. 2010. A Mean Field Model of Work Stealing in Large-scale Systems. *SIGMETRICS Perform. Eval. Rev.* 38, 1 (June 2010), 13–24. https://doi.org/10.1145/1811099.1811042
[11] N. Gast and B. Gaujal. 2012. Markov chains with discontinuous drifts have differential inclusion limits. *Performance Evaluation* 69, 12 (2012), 623–642.
[12] Itai Gurvich et al. 2014. Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *The Annals of Applied Probability* 24, 6 (2014), 2527–2559.
[13] Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python. (2001–). http://www.scipy.org/ [Online; accessed 2017-06-15].
[14] H. K. Khalil. 1996. *Nonlinear Systems* (2nd ed.). Prentice-Hall, Englewood Cliffs, NJ.
[15] Thomas G Kurtz. 1970. Solutions of Ordinary Differential Equations as Limits of Pure Jump Markov Processes. *Journal of Applied Probability* 7 (1970), 49–58.
[16] Thomas G Kurtz. 1978. Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and Their Applications* 6, 3 (1978), 223–240.
[17] G. Latouche and V. Ramaswami. 1999. *Introduction to Matrix Analytic Methods and stochastic modeling*. SIAM, Philadelphia.
[18] Jean-Yves Le Boudec, David McDonald, and Jochen Mundinger. 2007. A generic mean field convergence result for systems of interacting objects. In *Quantitative Evaluation of Systems, 2007. QEST 2007. Fourth International Conference on the*. IEEE, 3–18.
[19] M. Lin, B. Fan, J.C.S. Lui, and D. Chiu. 2007. Stochastic analysis of file-swarming systems. *Performance Evaluation* 64, 9 (2007), 856–875.
[20] L. Massoulié and M. Vojnović. 2005. Coupon Replication Systems. *SIGMETRICS Perform. Eval. Rev.* 33, 1 (June 2005), 2–13. https://doi.org/10.1145/1071690.1064215
[21] W. Minnebo and B. Van Houdt. 2014. A Fair Comparison of Pull and Push Strategies in Large Distributed Networks. *IEEE/ACM Transactions on Networking* 22 (2014), 996–1006. Issue 3.

[22] Michael Mitzenmacher. 2001. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems* 12, 10 (2001), 1094–1104.
[23] Michael David Mitzenmacher. 1996. *The Power of Two Random Choices in Randomized Load Balancing*. Ph.D. Dissertation. PhD thesis, Graduate Division of the University of California at Berkley.
[24] Cristopher Moore. 1990. Unpredictability and undecidability in dynamical systems. *Physical Review Letters* 64, 20 (1990), 2354.
[25] Vu Ngoc Phat and Tran Tin Kiet. 2002. On the Lyapunov equation in Banach spaces and applications to control problems. *International Journal of Mathematics and Mathematical Sciences* 29, 3 (2002), 155–166.
[26] V. Simoncini. 2016. Computational methods for linear matrix equations. *SIAM Rev.* 58 (2016), 377–441. Issue 3.
[27] John N Tsitsiklis and Kuang Xu. 2011. On the power of (even a little) centralization in distributed processing. *ACM SIGMETRICS Performance Evaluation Review* 39, 1 (2011), 121–132.
[28] Benny Van Houdt. 2013. A mean field model for a class of garbage collection algorithms in flash-based solid state drives. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 41. ACM, ACM, New York, NY, USA, 191–202.
[29] Nikita Dmitrievna Vvedenskaya, Roland L'vovich Dobrushin, and Fridrikh Izrailevich Karpelevich. 1996. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii* 32, 1 (1996), 20–34.
[30] Q. Xie, X. Dong, Y. Lu, and R. Srikant. 2015. Power of D Choices for Large-Scale Bin Packing: A Loss Model. *SIGMETRICS Perform. Eval. Rev.* 43, 1 (June 2015), 321–334. https://doi.org/10.1145/2796314.2745849
[31] Lei Ying. 2016. On the Approximation Error of Mean-Field Models. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science (SIGMETRICS '16)*. ACM, New York, NY, USA, 285–297. https://doi.org/10.1145/2896377.2901463
[32] Lei Ying. 2017. Stein's Method for Mean Field Approximations in Light and Heavy Traffic Regimes. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1, 1 (2017), 12.